FISTA Review

Yue Zhang

This review is based on the paper:
'A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems' by Beck and Teboulle 2008.
(* Questions are appreciated *)

June 18, 2015

Outline

Precursors: problem setup and lemmas

Algorithms: ISTA and FISTA

Recall ADMM

- We have the following constrained problem:

$$\begin{array}{ll} \text{minimize} & f(x) + g(z) \\ \text{subject to} & Ax + Bz = c \end{array}$$

- Augmented Lagrangian:

$$L_\rho(x, y) = f(x) + g(z) + y^T(Ax + Bz - c) + (\rho/2)\|Ax + Bz - c\|_2^2.$$

- Iterative:

$$\begin{aligned} x^{k+1} &:= \arg\min_x L_\rho(x, z^k, y^k) \\ z^{k+1} &:= \arg\min_z L_\rho(x^{k+1}, z, y^k) \\ y^{k+1} &:= y^k + \rho(Ax^{k+1} + Bz^{k+1} - c) \end{aligned}$$

- Note here A and B are not necessarily to be full rank.

Problem setup for FISTA

- For FISTA, we deal with the following problem:

$$\text{minimize} \quad f(x) + g(x)$$

  where $g : \mathbb{R}^n \to \mathbb{R}$ is continuous convex function and possibly nonsmooth. $f : \mathbb{R}^n \to \mathbb{R}$ is smooth convex function and continuous differentiable with Lipschitz constant $L(f)$:

$$\|\nabla f(x) - \nabla f(y)\| \leq L(f)\|x - y\|$$

  , People also say $\mu I \preccurlyeq \nabla^2 f(x) \preccurlyeq LI$.

- This is equivalent to:

$$\begin{array}{ll} \text{minimize} & f(x) + g(z) \\ \text{subject to} & x - z = 0. \end{array}$$

Revisit of Gradient Descent

- Solving minimize $f(x) + 0$, we have the update

$$x_k = x_{k-1} - t_k \nabla f(x_{k-1})$$

- It's said that it is **well known** this iteration can be viewed as a proximal regularization of linearized f at x_{k-1}:

$$x_k = \arg\min_x \{ f(x_{k-1}) + < x - x_{k-1}, \nabla f(x_{k-1}) > + \frac{1}{2t_k} \|x - x_{k-1}\|^2 \}$$

- Complete the square:

$$x_k = \arg\min_x \{ \frac{1}{2t_k} \|x - (x_{k-1} - t_k \nabla f(x_{k-1}))\|^2 \}$$

- If we have $g(x) = \lambda \|x\|_1$, this goes to (ISTA):

$$x_k = \arg\min_x \{ \frac{1}{2t_k} \|x - (x_{k-1} - t_k \nabla f(x_{k-1}))\|^2 + \lambda \|x\|_1 \}$$

Approximation Model

- Given $L > 0$ (this L is just a constant we select), we can approximate $F(x) := f(x) + g(x)$ by

$$Q_L(x, y) := f(y) + < x - y, \nabla f(y) > + \frac{L}{2}\|x - y\|^2 + g(x),$$

- It has a unique minimizer:

$$p_L(y) := \arg\min_x Q_L(x, y).$$

Same as before, we can get:

$$p_L(y) := \arg\min_x \{g(x) + \frac{L}{2}\|x - (y - \frac{1}{L}\nabla f(y))\|^2\}.$$

The ISTA step reduces to the follows:

$$x_k = p_L(x_{k-1})$$

- Note that if L is big, say $L > L(f)$,

$$f(x) \leq f(y) + < x - y, \nabla f(y) > + \frac{L}{2}\|x - y\|^2$$

Important Lemma

- Let $y \in \mathbb{R}^n$ and $L > 0$ be such that

  $$F(p_L(y)) \leq Q(p_L(y), y),$$

  Then for any $x \in \mathbb{R}^n$,

  $$F(x) - F(p_L(y)) \geq \frac{L}{2}\|p_L(y) - y\|^2 + L < y - x, p_L(y) - y >$$

- This guarantees that our target function is decreasing, thus can be taken as a backtracking checking criterion.
- From the former slide, if L is quite large, the assumption is guaranteed. However, $t_k = \frac{1}{L}$ can be small. Therefore we would like to choose L as small as possible which can still make the assumption satisfied. In other words, we need to approximate $L(f)$.

Quick proof

We have

$$F(x) - F(p_L(y)) \geq F(x) - Q(p_L(y), y),$$

Since f, g are convex, we have

$$f(x) \geq f(y) + < x - y, \nabla f(y) >$$
$$g(x) \geq g(p_L(y)) + < x - p_L(y), \partial g(y) >$$

Summing up these two inequalities,

$$F(x) \geq f(y) + < x - y, \nabla f(y) > + g(p_L(y)) + < x - p_L(y), \partial g(y) >$$

Recall

$$Q_L(p_L(y)) := f(y) + < p_L(y) - y, \nabla f(y) > + \frac{L}{2} \|p_L(y) - y\|^2 + g(p_L(y))$$

Note there is an implicit relation

$$\nabla f(y) + L(p_L(y) - y) + \partial g(y) = 0$$

Quick proof

Therefore,

$$\begin{aligned} F(x) - F(p_L(y)) &\geq F(x) - Q(p_L(y), y) \\ &\geq -\frac{L}{2}\|p_L(y) - y\|^2 + < x - p_L(y), \nabla f(y) + \partial g(y) > \\ &= -\frac{L}{2}\|p_L(y) - y\|^2 + L < x - p_L(y), y - p_L(y) > \\ &= -\frac{L}{2}\|p_L(y) - y\|^2 + L < p_L(y) - x, p_L(y) - y > \\ &= -\frac{L}{2}\|p_L(y) - y\|^2 + L < p_L(y) - y + y - x, p_L(y) - y > \\ &= \frac{L}{2}\|p_L(y) - y\|^2 + L < y - x, p_L(y) - y > \end{aligned}$$

Outline

Precursors: problem setup and lemmas

Algorithms: ISTA and FISTA

Two modes of ISTA

- **ISTA with constant stepsize**.
  Input: $L := L(f)$
  Step0. Take $x_0 \in \mathbb{R}^n$.
  Stepk. $(k \geq 1)$ Compute

$$x_k = p_L(x_{k-1})$$

.

- **ISTA with backtracking**.
  Step0. Take $L_0 > 0$, some η >1, and $x_0 \in \mathbb{R}^n$.
  Stepk. $(k \geq 1)$ Find the smallest nonnegative integer i_k such that with $\hat{L} = \eta^{i_k} L_{k-1}$,

$$F(p_{\hat{L}}(x_{k-1})) \leq Q_{\hat{L}}(p_{\hat{L}}(x_{k-1}), x_{k-1})$$

  Set $L_k = \eta^{i_k} L_{k-1}$ and compute

$$x_k = p_{L_k}(x_{k-1})$$

FISTA

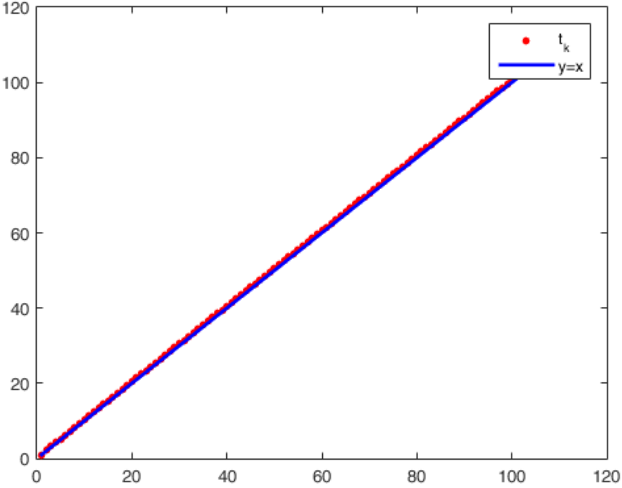
- **FISTA with backtracking stepsize**.
  Step0. Take $L_0 > 0$, some η >1,, $y_1 = x_0 \in \mathbb{R}^n$, $t_1 = 1$.
  Stepk. $(k \geq 1)$ Find the smallest nonnegative integer i_k such that with $\hat{L} = \eta^{i_k} L_{k-1}$,

$$F(p_{\hat{L}}(x_{k-1})) \leq Q_{\hat{L}}(p_{\hat{L}}(x_{k-1}), x_{k-1})$$

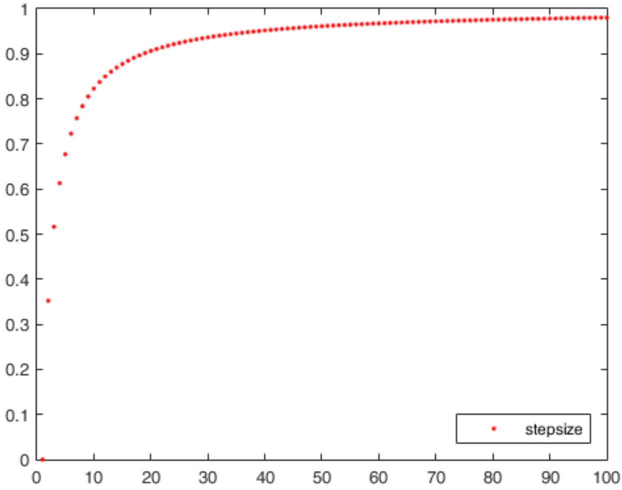
  Set $L_k = \eta^{i_k} L_{k-1}$ and compute

$$x_k = p_{L_k}(y_k)$$
$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}$$
$$y_{k+1} = x_k + (\frac{t_k - 1}{t_{k+1}})(x_k - x_{k-1})$$

About t_k



About Stepsize



Convergence Rate of ISTA and FISTA

If $\beta L(f) \leq L_k \leq \alpha L(f)$, which is guaranteed by our backtracking,

- For ISTA,

$$F(x_k) - F(x^*) \leq \frac{\alpha L(f)\|x_0 - x^*\|^2}{2k}$$

  Actually. it's

$$F(x_k) - F(x^*) \leq \frac{\alpha L(f)\|x_0 - x^*\|^2}{2k} - \frac{\alpha L(f)}{2k}\left(\frac{\beta}{\alpha}\sum_{n=0}^{k-1} n\|x_n - x_{n+1}\|^2 + \|x^* - x_k\|^2\right)$$

- For FISTA,

$$F(x_k) - F(x^*) \leq \frac{2\alpha L(f)\|x_0 - x^*\|^2}{(k+1)^2}$$

  (FISTA is faster than ADMM. In my opinion, generally this can be tricky.)

Speed Limit

Nesterov (2004) gives a simple example of a smooth function for which no method that generates iterates of the form $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$ can converge at a rate faster than $\frac{1}{k^2}$, at least for its first n/2 iterations.

$$A = \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \dots & 0 \\ 0 & & .. & .. & 0 \\ \vdots & & & .. & \vdots \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix}$$

$f(x) = 1/2 x^T A x - e_1^T x$. It can be shown? that

$$f(x_k) - f(x^*) \geq \frac{3L\|x_0 - x^*\|^2}{32(k+1)^2}$$

References

A. Beck and M. Teboulle.
A fast iterative shrinkage-thresholding algorithm for linear inverse problems.
2008.

S. Wright.
Optimization algorithms in machine learning.
2010.